2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Physics-Informed Machine Learning for Efficient Sim-to-Real Data Augmentation in Micro-Object Pose Estimation

Zongcai Tan*, Lan Wei* and Dandan Zhang

***Abstract*— Precise pose estimation of optical microrobots is essential for enabling high-precision object tracking and autonomous biological studies. However, current methods rely heavily on large, high-quality microscope image datasets, which are difficult and costly to acquire due to the complexity of microrobot fabrication and the labour-intensive labelling. Digital twin systems offer a promising path for sim-to-real data augmentation, yet existing techniques struggle to replicate complex optical microscopy phenomena, such as diffraction artifacts and depth-dependent imaging. This work proposes a novel physics-informed deep generative learning framework that, for the first time, integrates wave optics-based physical rendering and depth alignment into a generative adversarial network (GAN), to synthesise high-fidelity microscope images for microrobot pose estimation efficiently. Our method improves the structural similarity index (SSIM) by 35.6% compared to purely AI-driven methods, while maintaining real-time rendering speeds (0.022 s/frame). The pose estimator (CNN backbone) trained on our synthetic data achieves 93.9%/91.9% (pitch/roll) accuracy, just 5.0%/5.4% (pitch/roll) below that of an estimator trained exclusively on real data. Furthermore, our framework generalises to unseen poses, enabling data augmentation and robust pose estimation for novel microrobot configurations without additional training data.**

I. Introduction

Optical microscopy (OM) is commonly integrated with microrobotic systems for observing and characterizing micro/nano objects, providing essential visual feedback for precise three-dimensional (3D) localization and pose determination [1]. Accurate visual perception of optical microrobots is crucial for biomedical tasks at micro and nanoscales, such as targeted delivery, micromanipulation, and microassembly [2]–[5]. However, microscopic image degradation due to optical defocusing, diffraction, and background noise significantly hinders robust microrobot tracking and pose estimation [6]. Thus, effective vision-based methods for tracking and pose estimation are critical for enhancing microrobot perception and reliability in biomedical applications.

Due to high costs of micro/nano-fabrication and difficulties in precise out-of-plane pose control, obtaining large, diverse, and well-labelled real-world datasets is technically challenging and expensive [7]. The scarcity of high-quality annotated data severely restricts the performance of data-driven pose estimation methods [8]–[10]. Zhang *et al.* proposed a Generative Adversarial Network (GAN)-based augmentation to generate synthetic microrobot images to supplement real-world data [4]. However, their approach relies solely on learned data distributions, neglecting optical microscopy physics, causing discrepancies between synthetic and real images. Consequently, these images often miss critical optical phenomena necessary for high-fidelity microrobot imaging. Additionally, GANs trained on limited datasets may poorly generalize to unseen configurations, resulting in poor performance when the synthetic data does not fully represent the diversity of microrobot poses.

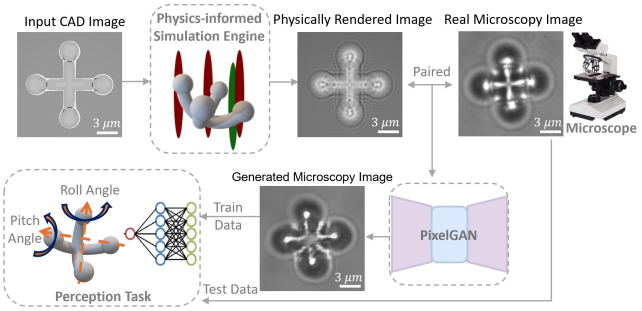

Fig. 1. Concept overview of the physics-informed machine learning network for efficient sim-to-real microscopy data generation.

Most existing physical simulation methods focus on fluorescence and electron microscopy, with limited research dedicated to modeling OM images [11]–[15]. Current simulation methods face a trade-off: either achieving high physical accuracy with substantial computational costs unsuitable for real-time applications, or simplifying physics, sacrificing generality and reliability [12], [13]. This constraint also hampers high-fidelity synthetic data generation needed to augment experimental datasets for training pose estimation models.

Here, this work proposes a high-fidelity digital twin system integrating physics-informed machine learning to facilitate sim-to-real OM image generation. Specifically, the system simulates key optical effects such as defocus blur, diffraction rings, and depth-dependent variations by applying wave optics principles, optimizing computational complexity for efficient simulation. Additionally, simulated images are aligned with experimental data using a pixel-to-pixel generative adversarial network (PixelGAN) [16], further reducing discrepancies between datasets. Image fidelity and efficiency are evaluated using the Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), and image generation inference time. Models trained on simulated and real datasets were tested on microrobot pose estimation tasks. Results confirm that integrating physics-based modeling with data-driven refinement

*Equal Contribution.

Zongcai Tan, Lan Wei and Dandan Zhang are with the Department of Bioengineering, Imperial-X Initiative, Imperial College London, London, United Kingdom. Corresponding: d.zhang17@imperial.ac.uk. Additional project details are available on our website. Code is available here.

improves both interpretability and realism, preserving crucial depth-encoding features and enhancing sim-to-real transfer.

In summary, we introduce the first physics-informed PixelGAN framework for developing a digital twin of an OT-actuated, complex-shaped microrobotic system. As illustrated in Fig. 1, the primary innovation is integrating physics-based rendering and pixel-level depth alignment into GAN training, significantly enhancing synthetic image realism and fidelity in depth encoding. This method effectively addresses data scarcity for microrobot perception and balances simulation fidelity with computational efficiency.

The **Main Contributions** of this paper are as follows:

1) The work introduces a high-fidelity digital twin system that integrates physics-informed machine learning to simulate OM images of optically actuated microrobots. By incorporating wave optics principles, the model accurately reproduces key optical effects, bridging the gap between simulation and real-world imaging.
2) To enhance simulation realism, this work employs PixelGAN, a deep generative model that refines simulated images to better align with real experimental data. This approach significantly improves dataset quality for visual perception algorithms, facilitating accurate and robust pose estimation of optical microrobots.
3) The system achieves real-time generation of high-fidelity microscopic images. By optimizing computational complexity without sacrificing fidelity, it accelerates training efficiency, increases dataset diversity, and reduces data collection and labelling costs, enabling broader microrobot applications in real-time visual tasks.
4) The proposed framework demonstrates generalisability to unseen pose configurations. PixelGAN-30 (five held-out poses) still yields data that trains high-accuracy estimators on unseen poses (only 2.4%-2.5% drop relative to PixelGAN-35), demonstrating robustness.

II. Related Work

A. Simulation of OM Imaging for Micro-Objects

In optical microscopy, noise and artifacts such as diffraction rings and defocus blur often encode critical information about an object's pose and depth. Even minor mismatches between simulated and real imaging can introduce significant errors in training visual algorithms. Current research primarily focuses on super-resolution reconstruction [11], with limited attention to the physical modeling of OM imaging. For example, Nasse and Woehl proposed a point spread function (PSF) model based on vector theory, but it struggles to handle fully 3D imaging of multilayered or complex 3D samples [17]. Marian's research on 3D amplitude PSF offers physically based image simulations, but its high computational demand limits real-time applications [18]. Wang et al. employed geometric optics ray tracing, Mie theory, and FDTD to simulate nano-scale microsphere imaging, but these methods face computational challenges, particularly with complex interference and wave effects [12]. Similarly, Zhang et al. introduced a diffractive optical element (DOE)-based approach combined with deep learning for image simulation and reconstruction. While this method simplifies component modeling to reduce complexity, it sacrifices accuracy in representing real optical pathways and assumes a uniform PSF over extended depth ranges, which limits its generalizability in complex samples [13]. Li et al. simulated confocal microscopy imaging focused on fluorescent signals but did not achieve virtual-real synchronization, neglecting continuous complex structures [14]. Overall, significant gaps remain in OM visualization and rendering, particularly in high-fidelity models that support fast and accurate imaging of complex 3D structures.

B. Machine Learning for Micro/Nanorobot Tracking

Supervised learning methods have been developed to achieve robust and precise micro-object tracking [19]. For example, deep neural network–based methods have been developed to estimate the 3D pose and depth of optically transparent microrobots [20]. Deep residual neural networks combined with Gaussian process regression have been shown to simultaneously estimate 3D position and orientation [21]. Moreover, a deep learning approach has been proposed to localize multiple microrobots and detect elliptical features, thereby enabling automated manipulation via OT [22], [23]. The performance of machine learning models for microrobot tracking and pose estimation is highly dependent on the quality of labeled datasets. However, the high costs of microrobot fabrication, data acquisition, and manual labeling have led to a severe shortage of such datasets. As a result, developing data augmentation techniques that can rapidly and cost-effectively generate large-scale, high-quality training data remains a major bottleneck for advancing machine learning-based tracking algorithms.

III. Methodology

A. System Overview and Workflow

To accurately replicate the imaging process in the OT system, this work develops a high-fidelity microscopy simulation model based on wave optics. The model incorporates the entire optical path of the microscope and accounts for key physical factors that influence image quality, including objective focal length ($f_{\text{obj}} = 50\,\text{mm}$), eyepiece focal length ($f_{\text{eye}} = 20\,\text{mm}$), numerical aperture (NA $= 1.45$), and illumination wavelength ($\lambda = 632.8\,\text{nm}$). These parameters govern resolution, magnification, and overall imaging fidelity. By coupling physics-based rendering with a GAN, the digital-twin images are refined to closely match experimental observations.

Fig. 2 summarizes the workflow of the visualization rendering system, coupling (i) virtual acquisition and preprocessing, (ii) depth-aware wave-optics rendering, and (iii) sim-to-real refinement via PixelGAN. Specifically, the pipeline begins by constructing a virtual optical system on the NVIDIA platform, generating an optical model from real-time optical path parameters and simulated robotic poses. A virtual camera then captures initial CAD images together

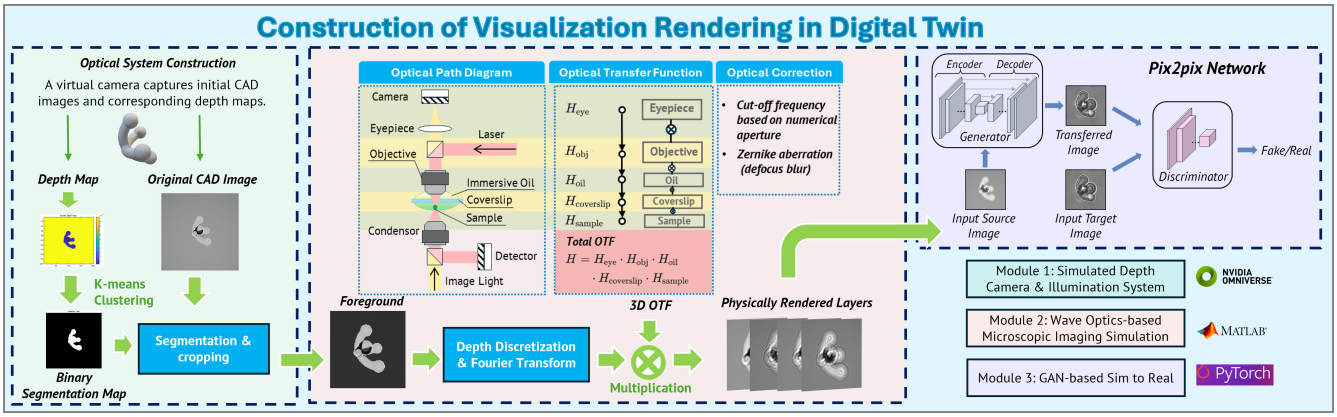

Fig. 2. Workflow of the visualization rendering system: A virtual optical microscope system was constructed in Isaac Sim based on real-time optical path parameters and predicted robotic poses. Using the initial CAD images and depth maps captured by a virtual camera, high-fidelity simulated images are generated via the visualization rendering module based on wave optics. The reality gap of the virtual image was further reduced through a sim-to-real module using PixelGAN [16].

with the corresponding depth maps, synchronized with real experimental conditions. The depth map is further processed via k-means clustering to segment the foreground (micro-robot) from the background, enabling tight cropping that encloses the robot and reduces computational load. Following foreground segmentation, image formation starts by deriving the microscope optical transfer function (OTF) from the wave-optics model of the optical path. The robot depth is subsequently discretized along the z-axis into multiple layers based on the depth map. Each depth layer is transformed into the Fourier domain and multiplied by its corresponding OTF, avoiding traditional spatial-domain convolution. During this frequency-domain processing, the NA cutoff is enforced by setting the OTF to zero for spatial frequencies exceeding the maximum resolvable frequency (f_{cutoff}), removing non-physical components and ensuring realistic optical limits. Parseval's theorem is applied throughout to preserve energy across Fourier operations. Finally, to bridge the simulation-to-reality gap, rendered images are refined by a sim-to-real module based on PixelGAN [16], reducing visual discrepancies and improving alignment with experimental data.

B. Optical System and Wave Optics Simulation

To simulate the microscope's optical system accurately, a detailed optical path model was constructed, as depicted in Fig. 2. This model encompasses critical optical components: objective lens, eyepiece, cover slip, immersion oil, and sample medium (deionized water). Incorporating precise optical properties, including the refractive indices of immersion oil ($n_{\text{oil}} = 1.515$), the coverslip ($n_{\text{coverslip}} = 1.515$), and the sample medium ($n_{\text{sample}} = 1.33$), is crucial. Variations in refractive index significantly affect the NA, diffraction patterns, and effective wavelength within different media, influencing the achievable resolution and imaging accuracy.

Fourier optics was employed to simulate light propagation accurately, with each optical component represented by an OTF that encapsulates its impact on the propagating wavefront. Key transfer functions for optical elements are expressed as follows:

- **Eyepiece OTF**:

$$H_{\text{eye}} = \exp\left(-i\pi \frac{(U^2+V^2)\lambda_e}{f_{\text{eye}}}\right) \tag{1}$$

- **Objective OTF**:

$$H_{\text{obj}} = \exp\left(-i\pi \frac{(U^2+V^2)\lambda_e}{f_{\text{obj}}}\right) \tag{2}$$

- **Cover Slip OTF**:

$$H_{\text{coverslip}} = \exp\left(i2\pi\lambda_{\text{coverslip}}(U^2+V^2)\right) \tag{3}$$

The combined effect of these components yields the total system OTF:

$$H_{\text{total}} = H_{\text{eye}} \times H_{\text{obj}} \times H_{\text{coverslip}} \times H_{\text{oil}} \times H_{\text{sample}} \tag{4}$$

The NA cutoff frequency, representing the highest spatial frequency resolvable by the optical system, is calculated as:

$$f_{\text{cutoff}} = \frac{\text{NA} \cdot n_{\text{oil}}}{\lambda} \tag{5}$$

Frequencies exceeding this limit are excluded to ensure the simulated image adheres strictly to the physical constraints of the microscope.

Optical aberrations due to lens imperfections were modeled using Zernike polynomials. Specifically, the fourth-order Zernike polynomial (Z_4), which describes primary spherical aberration (a common lens aberration causing blurred focal spots), was applied directly as an additional phase shift in the frequency domain:

$$H_{\text{total}}(U,V) \rightarrow H_{\text{total}}(U,V) \times \exp(iZ_4), \quad Z_4 = \sqrt{3}(2\rho^2-1) \tag{6}$$

where ρ is the normalized radial coordinate, defined as the radial distance from the optical axis normalized by the maximum radius set by the NA cutoff.

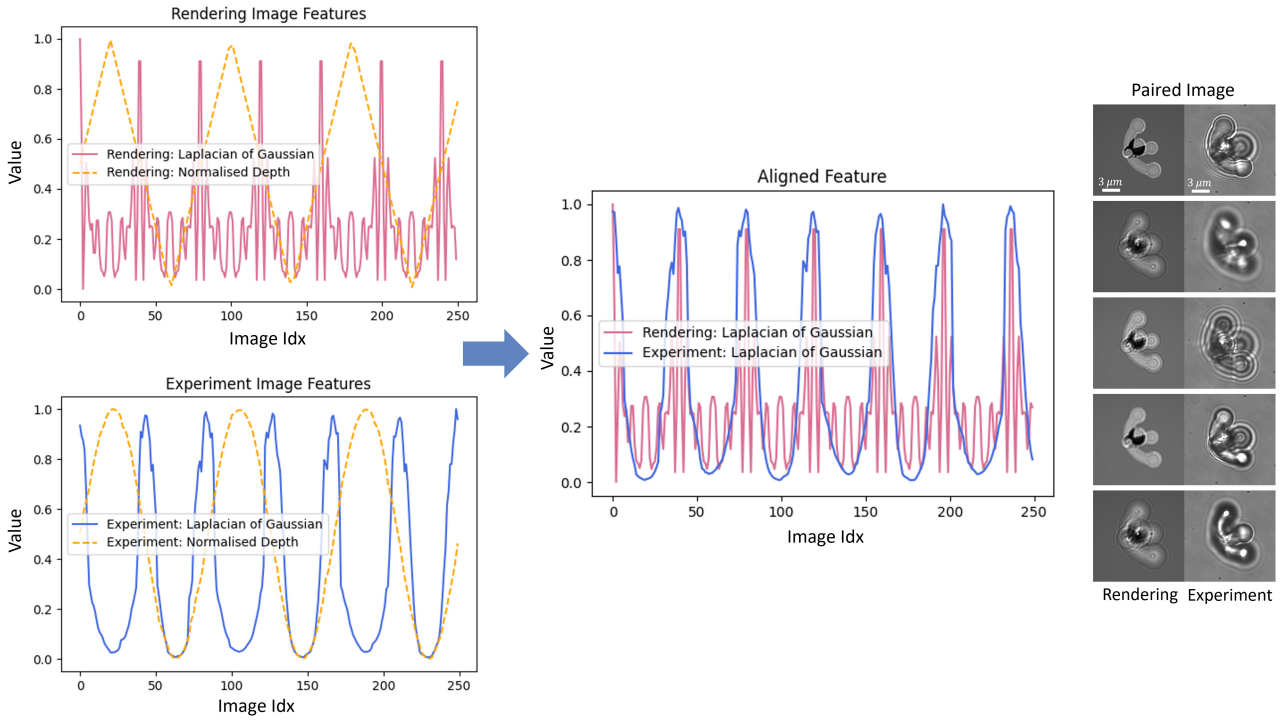

Fig. 3. Alignment of rendering and experiment image features based on Laplacian of Gaussian (LoG) analysis. LoG values and normalised depth are extracted for each dataset (left). Peak LoG frames (corresponding to the focal plane) are used to segment the datasets. To enable one-to-one pairing, data within each segment is balanced (middle), facilitating aligned image pairs for downstream training (right).

C. Depth Discretization and 3D Visual Rendering

As shown in Fig.2, to generate high-fidelity microscopic images, clear images and depth maps of the sample were obtained using Isaac Sim and transmitted to MATLAB via ROS. The microscope's focal plane was defined as the $0\,\mu\text{m}$ reference. The robot's operational depth, spanning from $-10\,\mu\text{m}$ to $+10\,\mu\text{m}$, was uniformly partitioned into 40 discrete layers, with each layer corresponding to a distinct robotic position and its respective optical projection. This discretization provides sufficiently fine axial sampling to capture depth-dependent optical effects while maintaining computational efficiency for real-time rendering. The angular spectrum propagation method was then used to calculate the OTF for each depth layer, and physically rendered images were generated by convolving the OTF with the corresponding depth robot image. Additionally, to improve real-time performance, parallel computing and GPU acceleration were employed, and Parseval's theorem was applied to ensure energy conservation across Fourier transformations.

D. Micro-Fabrication and Experimental Setting

The microrobots were fabricated using IP-L Photoresist (Nanoscribe, Germany) with a Nanoscribe 3D printer (Nanoscribe GmbH, Germany) through a two-photon polymerisation (2PP) process [24]. The microrobots used for collecting experimental data were printed on glass substrates and placed on a deionized water spacer (DI) [25]. The experimental setup consists of an OT (Elliot Scientific, UK) integrated with a nanopositioner (Mad City Labs Inc.). The collected dataset comprises microscope images recorded via a CCD camera, capturing the depth of microrobot and their poses. The pitch angle ('P') refers to the pose of the robot along the x-axis, while the roll angle ('R') refers to the pose along the y-axis. As 10° is set as the resolution of change angle, the study has a total of 35 different kinds of pose classes for the optical microrobot designed by Zhang et al. [25]. Each image frame has a resolution of 678×488 pixels. To obtain depth values, the printed microrobots were attached to a glass plane and positioned on a piezoelectric substrate, generating the z-axis trajectories.

E. Sim-to-Real Transfer

Although physics-based visualization rendering models can generate virtual microscope images that preserve depth information, they often exhibit discrepancies from real experimental images, particularly in terms of contrast and gloss. To address this, the work employs a PixelGAN further to reduce the gap between simulated and real images. This work first aligns features between physically rendered images and their corresponding experimental images at the same depths. Because the sharpness of an object's edges varies with depth under a microscope, edge detection becomes a key factor in this alignment. The Laplacian operator, as a second-order derivative, effectively identifies edges by capturing rapid changes in intensity [26], and combining it with Gaussian smoothing provides a robust feature for the alignment process.

First, the Laplacian of Gaussian (LoG) values are computed for each image in both datasets as shown in the left part of Fig.3. The image corresponding to the peak LoG

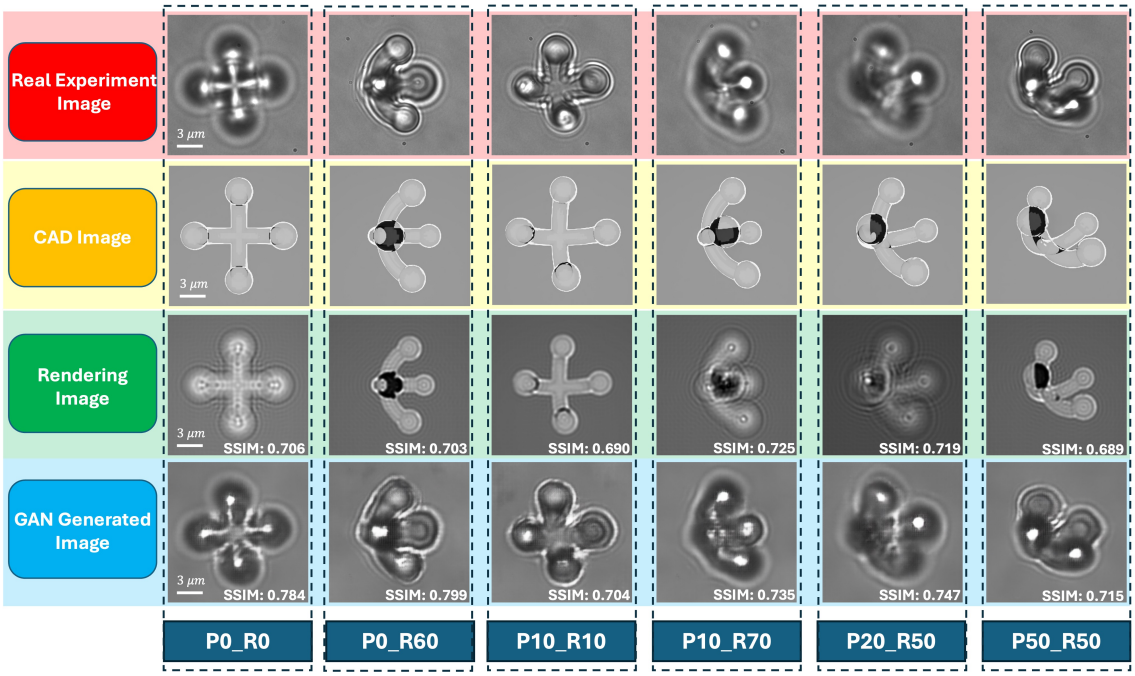

Fig. 4. Qualitative evaluation of image generation methods across varying poses and depths, demonstrating the visual fidelity of simulated microscope images compared to real experimental images. The comparison includes real experimental images (red), CAD renderings (yellow), physically rendered images (green), and GAN-generated images (blue).

value is identified as the frame located at the microscope's focal plane, where the normalised depth is 0, and the edges of the target object are the sharpest. Using the peak LoG points from the physically rendered images and experimental images, the datasets are divided into multiple segments. For each segment, the number of physically rendered images and experimental images is balanced by randomly selecting data points from the segment with more data. This ensures that the physically rendered images and experimental images have the same number of samples, allowing them to be paired one-to-one for the subsequent PixelGAN training.

PixelGAN consists of two main components: a Generator (G) and a Discriminator (D) [16]. It learns a mapping from the physics-rendered image x and a random noise vector z to the corresponding real experimental image y, formulated as: $\{x, z\} \rightarrow y$. The generator adopts a U-Net-style architecture, designed to synthesize realistic images that closely resemble real microscopy images. Meanwhile, the discriminator, implemented as a PatchGAN, operates adversarially to distinguish between real and generated images at the patch level, rather than evaluating entire images. This patch-based approach helps preserve fine-grained textures and structural consistency in the generated images. During training, the generator is optimized to fool the discriminator, while the discriminator simultaneously improves its ability to identify synthetic images, creating a competitive learning process. This training pipeline is illustrated in the right part of Fig. 2. The objective of the PixelGAN can be expressed as:

$$\begin{aligned}\mathcal{L}(G, D) =& \mathbb{E}_{x,y}[\log D(x, y)] + \\ & \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))],\end{aligned} \tag{7}$$

where G tries to minimize the objective against an adversarial D that tries to maximize it, i.e. $G^* = \arg\min_G \max_D \mathcal{L}(G, D)$. The work further adds an L1 distance to help the generator reduce blurring:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right]. \tag{8}$$

The final objective is:

$$G^* = \arg\min_G \max_D \mathcal{L}(G, D) + \lambda \mathcal{L}_{L1}(G). \tag{9}$$

In the implementation, the work sets λ as the default value 100, following the standard Pix2Pix configuration to balance adversarial realism and pixel-wise fidelity, which encourages structure-preserving translation critical for microscopy pose cues (e.g., edges and diffraction rings).

IV. Experiments and Results

A. Data

The aligned data used for PixelGAN training consists of 15,820 images (each consisting of one physically rendered image and one experiment image), corresponding to 35 sets of optical microrobots with different poses [27]. The resulting paired data are shown in the right part of Fig. 3, each pair has one rendered image and one experimental image on the same depth. Of these, 70% were allocated to the training set, 15% to the validation set, and 15% to the test set. The model is trained for 100 epochs. The code was implemented in PyTorch 1.8.1 and Python 3.8, running on a system equipped with 1 NVIDIA A100 GPU with 80 GB of memory. The CUDA version used was 11.4, and the inference precision was set to float32.

TABLE I. Performance comparison of different models. The best results in each column are shown in boldface.

Model/Value	Time (s) (↑)	SSIM (↑)	PSNR (↑)	MSE ($\times 10^{-2}$) (↓)
GAN	**0.002**	0.534	15.211	3.025
Rendering	0.020	0.639	14.728	3.587
Rendering + GAN	0.022	**0.724**	**18.370**	**1.548**

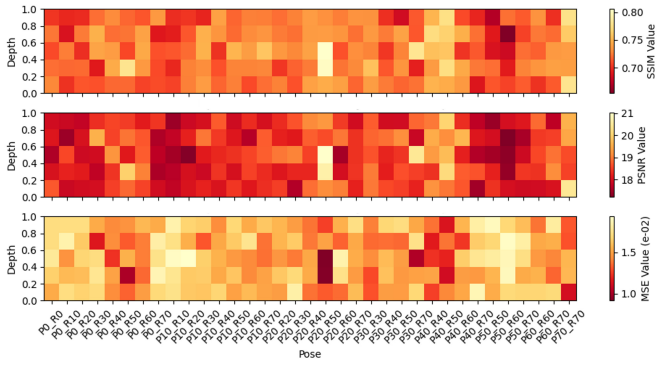

Fig. 5. Heatmap of evaluation metrics (SSIM, PSNR, MSE) across different robot poses and depths. The X-axis represents the robot's posture angles, while the Y-axis indicates the height offset relative to the focal plane. Each cell corresponds to a specific combination of pose and depth. The horizontal axis represents the robot's posture angles, denoted as `Pa_Rb`, where a and b indicate the pitch and roll angles in degrees, respectively (e.g., `P0_R60` means pitch = 0° and roll = 60°).

B. Evaluation Metrics

To evaluate the quality of the generated images, the work assessed both the image generation time and several quantitative metrics, including SSIM, PSNR, and MSE. MSE quantifies the average squared difference between pixel values of the ground-truth image ("target image") and the corresponding GAN-generated image. Lower MSE values indicate smaller pixel-wise differences and better alignment. PSNR, expressed in decibels (dB), measures the peak error between target and generated images. Higher PSNR values denote better image quality. SSIM evaluates structural similarity by analyzing luminance, contrast, and structural components. SSIM values range from 0 to 1, with values closer to 1 reflecting higher similarity and better visual quality. Inference of the GAN model was performed on a system equipped with an NVIDIA A100 GPU.

C. Image Generation Results

Table I compares three approaches for generating the images: 1) a GAN that uses only the CAD image as input, denoted as 'GAN'; 2) a physics-based rendering method using the CAD image, denoted as 'Rendering'; 3) a hybrid approach where a GAN takes the physics-based rendering as input, denoted as 'Rendering + GAN'. These comparisons serve as ablation studies.

The results show that combining physics-based rendering with the GAN significantly improves the overall image quality (by 35%) with only a minimal increase in per-image generation time (0.02 s). Fig. 4 provides example images from the real experiment, as well as corresponding CAD, rendered, and GAN-generated outputs, while Fig. 5 presents heatmaps illustrating the performance across each of the 35 pose classes. Notably, the GAN-generated images with physics-based rendering as input achieve the highest SSIM scores, surpassing those obtained from the physics-based rendering alone for each pose.

D. Pose Estimation

1) Result: The work further evaluates the quality of the generated data in the downstream task of microrobot pose estimation (pitch and roll), with results presented in Table II. Four evaluation metrics are used: accuracy, precision, recall, and F1 score. To investigate how network depth and inductive bias influence microscopy-image pose estimation, three backbones are selected: convolutional-based architectures (CNN and ResNet18) for their strong spatial feature extraction capabilities, and transformer-based model (ViT [28], which is good at capturing global dependencies. These architectures span different architectures, enabling a controlled study of (a) network depth (CNN vs ResNet-18) and (b) inductive bias (convolutional locality vs. global self-attention) under our limited-scale dataset.

Each model is trained or fine-tuned for 30 epochs to predict pitch and roll angles. The models are tested on the same experimental real data, consisting of 350 images representing 35 different poses. To avoid data leakage, which may inflate the pose estimation performance, the 350 test images were strictly excluded from the training sets of both the GAN and pose-estimation networks. Across three backbones, synthetic-data models underperform models trained on experimental data by 5.0%-18.8% in accuracy (pitch: 5.0%-16.0%; roll: 5.4%-18.8%). For the strongest CNN, the accuracy gaps are modest (pitch/roll: 5.0%/5.4%), demonstrating the high similarity between the generated and real datasets; for precision/recall/F1 the relative drops are 3.7%-4.0% (pitch) vs 6.2%-6.6% (roll). Notably, Table II shows that the shallow CNN performs best on generated images, likely because the physics-rendered + PixelGAN data are more regular and less diverse, so higher-capacity backbones (ResNet18/ViT) are more prone to overfitting residual sim-to-real artifacts, while a compact CNN better exploits locality-dominant cues (edges and diffraction rings). Moreover, ViT is typically more data-hungry, making its advantage harder to realize under our limited synthetic regime.

2) Hybrid Training: To evaluate the quality of our generated images, we conducted experiments using hybrid datasets for pose estimation model training with CNN backbone. We constructed multiple training datasets by combining experiment images with generated images at different ratios: 100% Exp, 75% Exp + 25% Gen, 50% Exp + 50% Gen, 25% Exp + 75% Gen, and 100% Gen, while maintaining consistent test evaluation on 350 experiment images across all experiments. The test images were strictly excluded from the training sets of both the GAN and pose-estimation networks. Each configuration was tested three times, and the results were averaged to ensure statistical reliability.

The experimental results, presented in Table III, demonstrate the effectiveness of our generated images in downstream pose estimation tasks.

TABLE II. Pose estimation results using models trained on experimental (*Exp*) and generated (*Gen*) images.

Model	Dataset	Accuracy (↑)		Precision (↑)		Recall (↑)		F1 Score (↑)	
		Pitch	Roll	Pitch	Roll	Pitch	Roll	Pitch	Roll
CNN	Exp	0.988	0.971	0.992	0.981	0.992	0.978	0.992	0.978
	Gen	0.939	0.919	0.955	0.916	0.954	0.916	0.952	0.914
ResNet18	Exp	**0.991**	**1.000**	**0.993**	**1.000**	**0.994**	**1.000**	**0.993**	**1.000**
	Gen	0.859	0.812	0.892	0.856	0.829	0.769	0.846	0.770
ViT	Exp	**0.991**	0.977	0.987	0.978	0.984	0.975	0.985	0.976
	Gen	0.832	0.812	0.844	0.819	0.757	0.779	0.772	0.781

Bold: best among Exp models; Underlined: best among Gen models.

TABLE III. Pose estimation results using CNN trained on hybrid experimental (Exp) and generated (Gen) images.

Hybrid Data	Accuracy (↑)		Precision (↑)		Recall (↑)		F1 Score (↑)	
	Pitch	Roll	Pitch	Roll	Pitch	Roll	Pitch	Roll
100% Exp	0.988	0.971	0.992	0.981	0.992	0.978	0.992	0.978
75% Exp + 25% Gen	0.981	0.949	0.985	0.962	0.975	0.960	0.980	0.959
50% Exp + 50% Gen	0.979	0.919	0.985	0.932	0.981	0.924	0.982	0.922
25% Exp + 75% Gen	0.959	0.921	0.966	0.919	0.955	0.949	0.965	0.951
100% Gen	0.939	0.919	0.955	0.916	0.954	0.916	0.952	0.914

TABLE IV. Pose Estimation Performance on Generated Images from PixelGAN Models Trained with Different Pose Sets. Average results over three experiments using the CNN-based pose estimation model. PixelGAN-35: trained on all 35 poses; PixelGAN-30: trained on Set B only (30 poses excluding Set A: P0_R20, P10_R30, P20_R40, P30_R50 and P40_R60). Both models generated images for all 35 poses for the pose estimation model training.

Model	Accuracy (↑)		Precision (↑)		Recall (↑)		F1 Score (↑)	
	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
PixelGAN-35	0.888	0.938	0.855	0.945	0.820	0.933	0.828	0.937
PixelGAN-30	0.866	0.916	0.707	0.915	0.661	0.896	0.675	0.899

Overall in hybrid training, accuracy decreases as the generated-image ratio increases: relative to 100% Exp, pitch drops by 0.7%-5.0% and roll by 2.3%-5.4%; precision/recall/F1 follow the same pattern (pitch ≈ 0.7%-4.0%, roll ≈1.9%-6.6%). Notably, replacing 50% of experimental images reduces pitch accuracy by only 0.9% versus 100% Exp, and roll accuracy at 100% Gen matches the 50% Gen case (both 0.919), indicating that modest mixing preserves accuracy.

3) Generalisability: To evaluate our model's capability in generating images for unseen poses, we conducted a generalisability experiment. We divided the 35 pose categories into Set A (P0_R20, P10_R30, P20_R40, P30_R50 and P40_R60) and Set B (the remaining 30 poses). Set A groups 5 challenging poses (hard to reach in physical experiments, structurally asymmetric, with strong optical artifacts), which makes them harder to acquire, to simulate faithfully, and to estimate accurately; this split provides a stringent generalisation test of PixelGAN on under-sampled, unseen cases. We trained two PixelGAN models under different conditions: 1) PixelGAN-35: trained on the complete dataset (Set A + Set B), and 2) PixelGAN-30: trained exclusively on Set B data. Subsequently, both models were used to generate images for all 35 pose categories, which were then employed to train pose estimation models.

The pose estimation models trained on generated images were evaluated on their classification performance across different pose sets. The results, presented in Table IV, demonstrate the generalisability. Set A is consistently harder than Set B across both generators, with accuracy lower by 5.3%-5.5%. Robustness to unseen poses is validated by comparing PixelGAN-30 with PixelGAN-35 on Set A (unseen for PixelGAN-30): accuracy shows a 2.5% relative drop (0.866 vs 0.888), which is smaller than the set's inherent difficulty. The accuracy difference between the two models on seen poses (Set B) is likewise small, at 2.4% (0.916 vs 0.938) Other metrics follow the same trend: on Set A, PixelGAN-30 vs PixelGAN-35 exhibits 0.148-0.159 absolute drops (precision/recall/F1), versus 3.2%-4.1% relative drops on Set B.

V. Discussion and Future Work

This study presents a physics-informed deep generative learning framework integrating wave optics-based physical modeling and depth alignment into PixelGAN, significantly enhancing sim-to-real data augmentation for microrobot pose estimation. Unlike purely data-driven or physics-based methods, this hybrid strategy employs accurate physical rendering to guide PixelGAN in efficiently capturing fine-grained visual features, such as depth-encoded diffraction rings and subtle optical artifacts, with minimal computational overhead. Specifically, the pixel-wise correspondence in PixelGAN preserves structural coherence and mid-level visual details crucial for accurate pose and depth estimation.

Model performance varied by data source: pre-trained ResNet18 achieved optimal results on experimental images, while a simpler 3-layer CNN excelled on generated images. This suggests that synthetic images maintain essential pose features but exhibit more consistent, less complex distributions than real experimental data. Consequently, simpler CNN architectures suit the relatively uniform synthetic data better, reducing the risk of overfitting compared to more complex models. The ViT showed suboptimal performance across all datasets, which can be attributed to its need for larger datasets to unlock its full potential.

Although this approach substantially reduces the sim-to-real performance gap and significantly lowers the cost

and complexity of experimental data collection, synthetic images still result in a minor performance difference (5.0%-5.4% with CNN pose estimator) compared to real-image-trained models. This residual gap arises partly from simplifications in the physics-based model and subtle variations in imaging conditions. Future research will address this by leveraging advanced domain adaptation techniques such as knowledge-guided transfer learning, feature-space alignment, and fine-tuning synthetic-trained models with limited experimental data. Further improvements to the physical simulation, including extending the current aberration modeling beyond primary spherical aberration to incorporate higher-order terms (e.g., astigmatism and coma), as well as modeling sensor-specific noise and fluidic diffraction effects, will further enhance realism and robustness.

VI. Conclusion

In summary, the work presents a physics-informed deep generative learning framework that significantly advances OM simulation for microrobot pose estimation. By combining wave-optics-based physical rendering with PixelGAN refinement, the method delivered a remarkable 35.6% improvement in image fidelity (SSIM) and achieved rapid data generation (0.022 seconds/frame), substantially narrowing the sim-to-real gap. The high accuracy (within 5.0%-5.4% for the CNN pose estimator) confirms the viability of synthetic data for training robust microrobot vision algorithms, greatly reducing reliance on labour-intensive experimental datasets. Additionally, the framework demonstrates generalisability to unseen pose configurations, with an $\sim$2.5% relative drop on unseen poses (accuracy), enabling robust deployment across diverse microrobotic scenarios. This work thus offers a practical and interpretable solution for efficient dataset augmentation, laying the foundation for safer and more explainable biomedical microrobotic systems.

References

[1] H. Dong, J. Lin, Y. Tao, Y. Jia, L. Sun, W. J. Li, and H. Sun, "Ai-enhanced biomedical micro/nanorobots in microfluidics," *Lab on a Chip*, vol. 24, no. 5, pp. 1419–1440, 2024.

[2] D. G. Grier, "A revolution in optical manipulation," *nature*, vol. 424, no. 6950, pp. 810–816, 2003.

[3] D. Zhang, Y. Ren, A. Barbot, F. Seichepine, B. Lo, Z.-C. Ma, and G.-Z. Yang, "Fabrication and optical manipulation of micro-robots for biomedical applications," *Matter*, vol. 5, no. 10, pp. 3135–3160, 2022.

[4] D. Zhang, A. Barbot, F. Seichepine, F. P.-W. Lo, W. Bai, G.-Z. Yang, and B. Lo, "Micro-object pose estimation with sim-to-real transfer learning using small dataset," *Communications Physics*, vol. 5, no. 1, p. 80, 2022.

[5] X. Sha, H. Sun, Y. Zhao, W. Li, and W. J. Li, "A review on microscopic visual servoing for micromanipulation systems: Applications in micromanufacturing, biological injection, and nanosensor assembly," *Micromachines*, vol. 10, no. 12, p. 843, 2019.

[6] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, "Reinforcement learning with artificial microswimmers," *Science Robotics*, vol. 6, no. 52, p. eabd9285, 2021.

[7] H. Li, Z. Zhang, X. Yi, S. Jin, and Y. Chen, "Control of self-winding microrobot using an electromagnetic drive system: integration of movable electromagnetic coil and permanent magnet," *Micromachines*, vol. 15, no. 4, p. 438, 2024.

[8] L. Yang, J. Jiang, F. Ji, Y. Li, K.-L. Yung, A. Ferreira, and L. Zhang, "Machine learning for micro-and nanorobots," *Nature Machine Intelligence*, vol. 6, no. 6, pp. 605–618, 2024.

[9] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *PeerJ Computer Science*, vol. 8, p. e1045, 2022.

[10] A. J. Plompen, O. Cabellos, C. de Saint Jean, M. Fleming, A. Algora, M. Angelone, P. Archier, E. Bauge, O. Bersillon, A. Blokhin *et al.*, "The joint evaluated fission and fusion nuclear data library, jeff-3.3," *The European Physical Journal A*, vol. 56, pp. 1–108, 2020.

[11] E. T. Rogers, J. Lindberg, T. Roy, S. Savo, J. E. Chad, M. R. Dennis, and N. I. Zheludev, "A super-oscillatory lens optical microscope for subwavelength imaging," *Nature materials*, vol. 11, no. 5, pp. 432–435, 2012.

[12] Z. Wang, W. Guo, L. Li, B. Luk'Yanchuk, A. Khan, Z. Liu, Z. Chen, and M. Hong, "Optical virtual imaging at 50 nm lateral resolution with a white-light nanoscope," *Nature communications*, vol. 2, no. 1, p. 218, 2011.

[13] Y. Zhang, X. Song, J. Xie, J. Hu, J. Chen, X. Li, H. Zhang, Q. Zhou, L. Yuan, C. Kong *et al.*, "Large depth-of-field ultra-compact microscope by progressive optimization and deep learning," *Nature Communications*, vol. 14, no. 1, p. 4118, 2023.

[14] Y. Li and F. Huang, "A statistical resolution measure of fluorescence microscopy with finite photons," *Nature Communications*, vol. 15, no. 1, p. 3760, 2024.

[15] D. Balakrishnan, S. W. Chee, Z. Baraissov, M. Bosman, U. Mirsaidov, and N. D. Loh, "Single-shot, coherent, pop-out 3d metrology," *Communications Physics*, vol. 6, no. 1, p. 321, 2023.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] M. J. Nasse and J. C. Woehl, "Realistic modeling of the illumination point spread function in confocal scanning optical microscopy," *Josa a*, vol. 27, no. 2, pp. 295–302, 2010.

[18] A. Marian, F. Charrière, T. Colomb, F. Montfort, J. Kühn, P. Marquet, and C. Depeursinge, "On the complex three-dimensional amplitude point spread function of lenses and microscope objectives: theoretical aspects, simulations and measurements by digital holography," *Journal of microscopy*, vol. 225, no. 2, pp. 156–169, 2007.

[19] Z. Cenev, J. Venäläinen, V. Sariola, and Q. Zhou, "Object tracking in robotic micromanipulation by supervised ensemble learning classifier," in *2016 International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS)*. IEEE, 2016, pp. 1–5.

[20] M. Grammatikopoulou and G.-Z. Yang, "Three-dimensional pose estimation of optically transparent microrobots," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 72–79, 2019.

[21] D. Zhang, F. P.-W. Lo, J.-Q. Zheng, W. Bai, G.-Z. Yang, and B. Lo, "Data-driven microscopic pose and depth estimation for optical microrobot manipulation," *Acs Photonics*, vol. 7, no. 11, pp. 3003–3014, 2020.

[22] Z. Tan and D. Zhang, "Interactive ot gym: A reinforcement learning-based interactive optical tweezer (ot)-driven microrobotics simulation platform," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 1–7.

[23] Y. Ren, M. Keshavarz, S. Anastasova, G. Hatami, B. Lo, and D. Zhang, "Machine learning-based real-time localization and automatic trapping of multiple microrobots in optical tweezer," in *2022 international conference on manipulation, automation and robotics at small scales (MARSS)*. IEEE, 2022, pp. 1–6.

[24] S. Kawata, H.-B. Sun, T. Tanaka, and K. Takada, "Finer features for functional microdevices," *Nature*, vol. 412, no. 6848, pp. 697–698, 2001.

[25] D. Zhang, A. Barbot, B. Lo, and G.-Z. Yang, "Distributed force control for microrobot manipulation via planar multi-spot optical tweezer," *Advanced Optical Materials*, vol. 8, no. 21, p. 2000543, 2020.

[26] X. Wang, "Laplacian operator-based edge detectors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 5, pp. 886–890, 2007.

[27] L. Wei and D. Zhang, "A dataset and benchmarks for deep learning-based optical microrobot pose and depth perception," *arXiv preprint arXiv:2505.18303*, 2025.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.